CLAIMS

What is claimed is:

1. A processor-implemented method for allocating resources to a plurality of applications, comprising:

gathering instrumentation data for work requests processed by the applications;

determining an associated workload level for work requests processed by the applications;

determining for each application a first application resource requirement as a function of the workload levels and a service level metric associated with the application;

determining for each application an assigned subset of resources as a function of the first application resource requirement, wherein the function minimizes communication delays between resources, and satisfies a bandwidth capacity requirement of the application; and

automatically reconfiguring the resources consistent with the assigned subset of resources for each application.

2. The method of claim 1, further comprising:

classifying the work requests by type of requester and type of work;

determining an associated requester-load level for each type of requester;

determining an associated workload level for each type of work; and

adjusting a load balancing policy as a function of the workload levels and requester-load level, wherein work requests are assigned to the resources according to the load balancing policy.

3. The method of claim 1, wherein the resources include a plurality of servers and at least one of the applications uses a tiered arrangement of servers, and the step of determining application resource requirements further comprises:

representing each server as a processor-sharing queue having at least one critical resource;

determining an approximate average response time for a selected number of servers in each tier as a function of each processor-sharing queue; and determining a minimum total number of servers required in each tier for an average response time of the application to satisfy the service level metric.

4. The method of claim 1, wherein at least one application uses a tiered arrangement of servers, the application has resource requirements associated with each tier, and the step of determining an assigned subset of resources comprises assigning resources to tiers by a function that satisfies the resource requirements associated with each tier and minimizes communication delay between servers.

5. The method of claim 4, wherein the function is a mixed-integer programming function.

6. The method of claim 4, wherein the step of determining an assigned subset of resources comprises:

determining an initial assignment of the subset of resources using a first mixed-integer programming function;

determining a feasible assignment of the subset of resources from the initial assignment using a non-linear programming function; and

determining a final assignment of the subset of resources from the feasible assignment using a second mixed-integer programming function.

7. A processor-implemented method for allocating resources to a plurality of applications, comprising:

storing work-request identifier data when a work request is initiated;

determining an identity of a completed work request from the work-request identifier data when a work request is complete and storing instrumentation data for identified work requests processed by the applications;

classifying the work requests by type of requester and type of work;

determining an associated requester-load level for each type of requester;

determining an associated workload level for each type of work for work requests processed by the applications;

adjusting a load balancing policy as a function of the workload levels and requester-load level, wherein work requests are assigned to the resources according to the load balancing policy;

generating for each application a first application resource requirement as a function of the workload levels and a service level metric associated with the application;

determining for each application an assigned subset of resources as a function of the first application resource requirement, wherein the function minimizes communication delays between resources, and satisfies a bandwidth capacity requirement of the application; and

automatically reconfiguring the resources consistent with the assigned subset of resources for each application.

8. The method of claim 7, wherein the resources include a plurality of servers and at least one of the applications uses a tiered arrangement of servers, and the step of determining application resource requirements further comprises:

representing each server as a processor-sharing queue having at least one critical resource;

determining an approximate average response time for a selected number of servers in each tier as a function of each processor-sharing queue;

determining a minimum total number of servers required in each tier for an average response time of the application to satisfy the service level metric.

9. The method of claim 7, wherein at least one application uses a tiered arrangement of servers, the application has resource requirements associated with each tier, and the step of determining an assigned subset of resources comprises assigning resources to tiers by a function that satisfies the resource requirements associated with each tier and minimizes communication delay between servers.

10. The method of claim 9, wherein the function is a mixed-integer programming function.

11. The method of claim 9, wherein the step of determining an assigned subset of resources comprises:

determining an initial assignment of the subset of resources using a first mixed-integer programming function;

determining a feasible assignment of the subset of resources from the initial assignment using a non-linear programming function; and

determining a final assignment of the subset of resources from the feasible assignment using a second mixed-integer programming function.

12. An apparatus for allocating resources to a plurality of applications, comprising:

means for gathering instrumentation data for work requests processed by the applications;

means for determining an associated workload level for work requests processed by the applications;

means for generating for each application a first application resource requirement as a function of the workload levels and a service level metric associated with the application;

means for determining for each application an assigned subset of resources as a function of the first application resource requirement, wherein the function minimizes communication delays between resources, and satisfies a bandwidth capacity requirement of the application; and

means for automatically reconfiguring the resources consistent with the assigned subset of resources for each application.

13. The apparatus of claim 12, further comprising:

means for classifying the work requests by type of requester and type of work;

means for determining an associated requester-load level for each type of requester;

means for determining an associated workload level for each type of work; and

means for adjusting a load balancing policy as a function of the workload levels and requester-load level, wherein work requests are assigned to the resources according to the load balancing policy.

14. The apparatus of claim 12, further comprising:

means for storing work-request identifier data when a work request is initiated; and means for determining an identity of a completed work request from the work-request identifier data when a work request is complete and storing instrumentation data for identified work requests processed by the applications.

15. An article of manufacture for allocating resources to a plurality of applications, comprising:

a computer-readable medium configured with instructions for causing a processor-based system to perform the steps of,

gathering instrumentation data for work requests processed by the applications;

determining an associated workload level for work requests processed by the applications;

generating for each application a first application resource requirement as a function of the workload levels and a service level metric associated with the application;

determining for each application an assigned subset of resources as a function of the first application resource requirement, wherein the function minimizes communication delays between resources, and satisfies a bandwidth capacity requirement of the application; and

automatically reconfiguring the resources consistent with the assigned subset of resources for each application.

16. The article of manufacture of claim 15, wherein the computer-readable medium is further configured with instructions for causing a processor-based system to perform the steps of:

classifying the work requests by type of requester and type of work;

determining an associated requester-load level for each type of requester;

determining an associated workload level for each type of work; and

adjusting a load balancing policy as a function of the workload levels and requester-load level, wherein work requests are assigned to the resources the according to the load balancing policy.

17. The article of manufacture of claim 15, wherein the resources include a plurality of servers and at least one of the applications uses a tiered arrangement of servers, and the computer-readable medium is further configured with instructions for causing a processor-based system to, in determining application resource requirements, perform the steps:

representing each server as a processor-sharing queue having at least one critical resource;

determining an approximate average response time for a selected number of servers in each tier as a function of each processor-sharing queue; and

determining a minimum total number of servers required in each tier for an average response time of the application to satisfy the service level metric.

18. The article of manufacture of claim 15, wherein at least one application uses a tiered arrangement of servers, the application has resource requirements associated with each tier, and the computer-readable medium is further configured with instructions for causing a processor-based system to, in determining an assigned subset of resources, perform the step of assigning resources to tiers by a function that satisfies the resource requirements associated with each tier and minimizes communication delay between servers.

19. The article of manufacture of claim 18, wherein the function is a mixed-integer programming function.

20. The article of manufacture of claim 18, wherein the computer-readable medium is further configured with instructions for causing a processor-based system to, in determining an assigned subset of resources, perform the steps of:

determining an initial assignment of the subset of resources using a first mixed-integer programming function;

determining a feasible assignment of the subset of resources from the initial assignment using a non-linear programming function; and

determining a final assignment of the subset of resources from the feasible assignment using a second mixed-integer programming function.

21. An article of manufacture for allocating resources to a plurality of applications, comprising:

a computer-readable medium configured with instructions for causing a processor-based system to perform the steps of,

storing work-request identifier data when a work request is initiated;

determining an identity of a completed work request from the work-request identifier data when a work request is complete and storing instrumentation data for identified work requests processed by the applications;

classifying the work requests by type of requester and type of work;

determining an associated requester-load level for each type of requester;

determining an associated workload level for each type of work for work requests processed by the applications;

adjusting a load balancing policy as a function of the workload levels and requester-load level, wherein work requests are assigned to the resources according to the load balancing policy;

generating for each application a first application resource requirement as a function of the workload levels and a service level metric associated with the application;

determining for each application an assigned subset of resources as a function of the first application resource requirement, wherein the function minimizes communication delays between resources, and satisfies a bandwidth capacity requirement of the application; and

automatically reconfiguring the resources consistent with the assigned subset of resources for each application.

22. The article of manufacture of claim 21, wherein the resources include a plurality of servers and at least one of the applications uses a tiered arrangement of servers, and the computer-readable medium is further configured with instructions for causing a processor-based system to, in determining application resource requirements, perform the steps:

representing each server as a processor-sharing queue having at least one critical resource;

determining an approximate average response time for a selected number of servers in each tier as a function of each processor-sharing queue;

determining a minimum total number of servers required in each tier for an average response time of the application to satisfy the service level metric.

23. The article of manufacture of claim 21, wherein at least one application uses a tiered arrangement of servers, the application has resource requirements associated with each tier, and the computer-readable medium is further configured with instructions for causing a processor-based system to, in determining an assigned subset of resources, perform the step of assigning resources to tiers by a function that satisfies the resource requirements associated with each tier and minimizes communication delay between servers.

24. The article of manufacture of claim 23, wherein the function is a mixed-integer programming function.

25. The article of manufacture of claim 23, wherein the computer-readable medium is further configured with instructions for causing a processor-based system to, in determining an assigned subset of resources, perform the steps of:

determining an initial assignment of the subset of resources using a first mixed-integer programming function;

determining a feasible assignment of the subset of resources from the initial assignment using a non-linear programming function; and

determining a final assignment of the subset of resources from the feasible assignment using a second mixed-integer programming function.